Using the scene to calibrate the camera

Tiago Trocoli, Luciano Oliveira
Federal University of Bahia (UFBA)
iVision Lab
Salvador, Brazil
http://ivisionlab.dcc.ufba.br

***Abstract*—Surveillance cameras are used in public and private security systems. Typical systems may contain a large number of different cameras, which are installed in different locations. Manual calibration of each single camera in the network becomes an exhausting task. Although we can find methods that semi-automatically calibrate a static camera, to the best of our knowledge, there is not a fully automatic calibration procedure, so far. To fill this gap, we propose here a novel framework for completely auto-calibration of static surveillance cameras, based on information of the scene (environment and walkers). Characteristics of the method include robustness to walkers' pose and to camera location (pitch, roll, yaw and height), and rapid camera parameter convergence. For a thorough evaluation of the proposed method, the walkers' foot-head projection, the length of the lines projected on the ground plane and the walkers' real heights were analyzed over public and private data sets, demonstrating the potential of the proposed method.**

***Keywords*-camera calibration; surveillance camera; auto calibration;**

I. Introduction

Camera calibration allows mapping of world coordinates into images, offering several benefits to other Computer Vision fields. For example, it is possible to decrease the computational effort of image object search [1], object tracking and pose estimation can be improved by 3D scene mapping [2], and 3D information of the scene can provide contextual information for people re-identification systems [3].

Research in camera auto-calibration becomes of paramount importance as the use of surveillance camera grows. Existing methods commonly require constrained assumptions of the scene to have a camera calibrated. Lv et al. [4] used head-foot homology to calibrate a camera, demanding a controlled scene with a tracked and previously known pedestrian in the scene. Krahnstoever et al. [5] addressed the camera auto-calibration problem by limited pedestrians detection, and then applying a bayesian network to find the necessary camera calibration parameters. Lv et al. [6] extracted two line segments from the ground plane, specially indicated by the user to detect vanishing points, in a controlled scene; a person reference must be selected only when legs cross during human walk in order to minimize the error in the measurement of a walker's major axis. Patwardhan et al. [7] applied a person detection to find the walkers' major axes; this constraint requires walkers in vertical orientation, in the scene. In a controlled indoor scene, Micusik et al. [8] used shape object detection to estimate walkers' major axes and the distance between objects and the camera; that turns the process to have an intensive computational effort, limiting the possible human poses to be searched in the images. Liu et al. [9] used a walker's major axis line, extracted by a background subtraction method, to calibrate a surveillance camera via a Bayesian approach; a large set of samples was required to achieve reasonable precision. Lee et al. [10] exploited scene clues for detection of vanishing points, requiring an object size as reference to

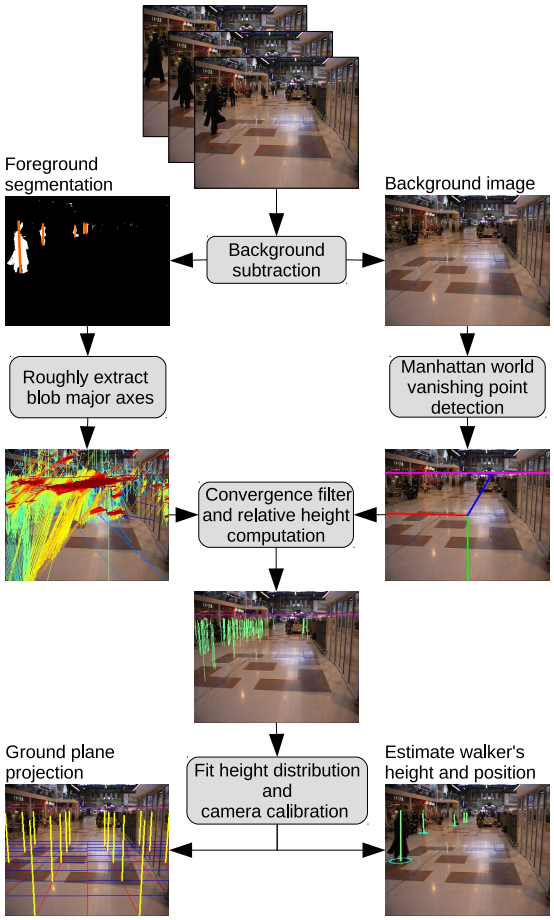

Fig. 1. Outline of the proposed auto-calibration framework. Given a sequence of images, an adaptive background subtraction extracts the background and walkers' blobs. Manhattan World based vanishing points are detected in the background image, and then used in a convergence filter to select the blobs, which point to vertical vanishing point. Walker's relative height are computed to fit a height distribution, and, finally, the required camera calibration parameters are estimated.

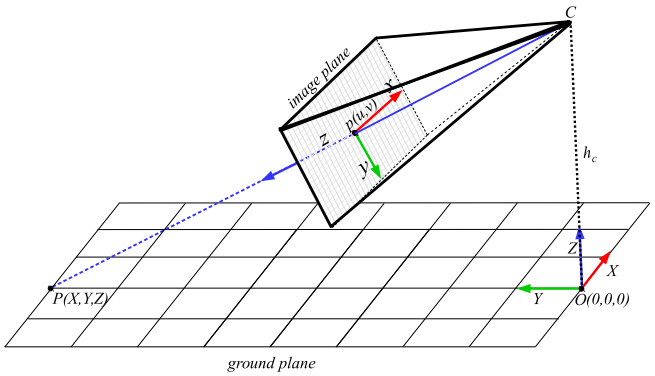

Fig. 2. Camera model and representation of world coordinate system.

work and a high level of human intervention.

A. Contributions

This paper introduces a method which can overcome the main limitations of the existing approaches, mainly regarding the constraints demanded for the other methods. Basic feature extraction from walkers and scenario structures is made fully automatic, not demanding any human intervention, which is required in almost all related works [4], [10]. Our method is robust to variation of walkers' pose, and camera position and orientation (pitch, roll, yaw and height), avoiding the use of person detectors (and, consequently, avoid failing when detection fails), as found in [7], [8], [5], [6]; this leads our method to have more flexibility and accuracy. The extraction of environment structure cues helps reducing the amount of samples necessary for camera calibration, in contrast to [9], which does not exploit this type of information. Table I summarizes the comparison at a glance.

The main goal of our method is to calibrate a pinhole-based camera fully automatic (see Sec. II). The main assumptions of the proposed method are twofold: (i) based on the Manhattan World constraint and (ii) considering prior information on walkers, walkers' major axes are first computed, and thereafter filtered by a vertical vanishing point convergence filter – as described in Sec. III. Scene structure features, computed against the image background, support vanishing point detection, which is addressed in Sec. IV. Relative height and height distribution are computed to fit the average of a walkers' relative height of the local distribution – described in Sec. V. After defining all required calibration parameters, the performance of the proposed method is assessed in Sec. VI. Sec VII draws the conclusions. All these main steps of the auto-calibration framework are depicted in Fig. 1.

II. Camera Model

Here the camera model is assumed to be pinhole. This assumption can be used for all types of static surveillance camera, without loss of generality. Thus, the relationship between the 3D point $P = [X, Y, Z, 1]^T$ and its image projection $p = [u, v, 1]^T$ is given by

$$p = K[R|t]P \tag{1}$$

where K is the intrinsic matrix, and $[R|t]$ represents the extrinsic matrix, composed of a rotation matrix $R = R_Z R_X R_Y$ and a translation vector $t = [0, 0, -h_c]^T$. As in [11], we assume zero-skew, aspect ratio equal to one and image principal point center $(0, 0)$, so that the focal length f is the only parameter to be found in K. In the world coordinate system (WCS), the origin point is placed on the ground plane as a projection of the camera center point. These assumptions narrow the camera calibration problem down to three required parameters (θ, ρ, h_c), which are needed besides f, to map a point in WCS to the camera coordinate system; θ and ρ represent the camera rotation around the X-axis ($R_X(\theta)$) and the Z-axis ($R_Z(\rho)$) in matrix R, respectively, and h_c denotes the translation (t) in Z-axis orientation. θ is bounded by $\pi/2$ (camera is parallel to the ground plane) and π (camera is pointing directly to ground plane). The yaw camera orientation does not contribute to camera calibration when the ground plane is taken as the reference. This assumption allows to set R_Y as a $3X3$ identity matrix [4].

Liebowitz and Zisserman [11] show that tree orthogonal vanishing points are directly associated with K and R so that it is possible to recover f, θ and ρ from only image information. So, by representing the vanishing points in image coordinates, as $vp_i = (u_i, v_i)$, where vp_0 is the vertical vanishing point, and vp_1 and vp_2 are the vanishing points of the horizon lines, f can be calculated as

$$f = \sqrt{d(center, vp_0) * l} \tag{2}$$

where $d(.,.)$ is the distance between two points, and l is the distance between the camera center point and the horizon line; now, θ and ρ can be defined as

$$\theta = atan2\left(\sqrt{u_0^2 + v_0^2}, -f\right) \tag{3}$$

$$\rho = atan(-u_0/v_0) \tag{4}$$

The last parameter to be found is the camera height, h_c. For that, the walkers' relative heights, h_i, achieved by the ratio of walkers' real height, h_i^{3D}, and h_c, are defined as

$$h_i = \frac{h_i^{3D}}{h_c} = 1 - \frac{d(p_h, q_l)d(p_f, vp_0)}{d(p_f, q_l)d(p_h, vp_0)} \tag{5}$$

where p_h and p_f are the highest and the lowest point, respectively, of walkers' major axes, and q_l is the intersection point between the horizon line and the line defined by p_h and p_f.

According to Criminisi, Reid and Zisserman [12], Eq. (5) provides a way to determine h_i by means of the horizon line and the vertical vanishing point vp_0. Each walkers' relative height keeps its value regardless of walker's major axis variation (e.g., due to perspective projection). This way, it is possible to apply the prior distribution of the walkers' real heights, h_i^{3D}, on h_i, with no need of knowing the real height of any walker in the monitored scene. By rewriting Eq. (5), it is possible to estimate h_c as the ratio between the average of the walkers' real height, H^{3D}, and the expectation, E, of h_i, given by

TABLE I
COMPARATIVE SUMMARY OF THE WORKS ON CAMERA CALIBRATION.

Ref	Walkers Features		Environment Features		Reference Measure	
	Manual	**Auto**	**Manual**	**Auto**	**Manual**	**Auto**
[4]	Background subtraction with manual axis				Track a walker with known height	
[5]		Limited walker detection, and specific to the scene			Require exactly average of walkers height	
[6]		Background subtraction with cross leg detection			Track a walker with known height	
[7]		Apply a walker detector with camera position constraints			Track a walker with known height	
[8]		Use specific shape person detector for camera position constraint			Know the person within controlled environment	
[9]		Background subtraction with main axis estimator				Uses the human height distribution as reference
[10]	Walker main axis inserted manually		4 lines marked on background scene, and each pair should point for differents vanishing points		Require a known height object	
Ours		**Based on completely automatic background subtraction**		**Completely automatic vanishing point detection, and scene line detection**		**Use height distribution and scene cues as reference**

$$h_c = \frac{H^{3D}}{E(h_i)} \tag{6}$$

Figure 2 illustrates the geometric representation of the camera model and its parameters.

III. BACKGROUND SUBTRACTION

Methods which use specialized person detection to estimate camera parameters limit the calibration process to a particular camera position and orientation [7], [8]. Instead, the goal here is to use background subtraction (BS) to provide a flexible way of extracting the walkers' major axes, in an unsupervised manner. This brings the advantage of not restricting the calibration method to a previously known scene. On the other hand, if the calibration method solely relies on blob extraction for estimation of camera parameters, this can fail if BS fails. Warped blobs tend to induce erroneous major axis orientation and size[1]. This is because the major axes do not intersect the vertical vanishing point anywhere (see Fig. 5(a)). To circumvent this orientation problem, after estimating the feet-head homology, ($\overline{p_f p_h}$), and the vanishing points (see Sec. IV), a convergence filter is applied in the major axes, achieved by the foreground blobs, in order to select those major axes that point to vp_0 (see Fig. 5(b)). To overcome the size problem, a RANSAC is used to select the major axes, which fit a height distribution; see Sec. V and Fig. 5(c), for more details.

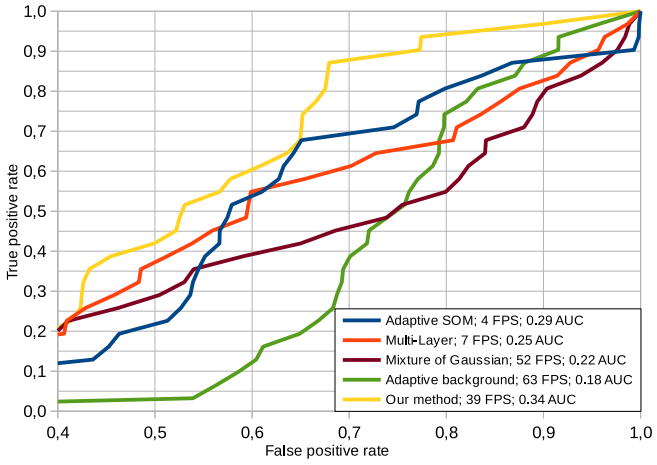

Fig. 3. Comparative performance of background subtraction methods.

Broadly speaking, the characteristics of a BS method to

[1]Here, blob shape is approximated by an ellipsis.

Fig. 4. Example of dilate and erode operators applied on walker blob.

accomplish our task should include speed and segmentation accuracy. Here we propose a modified version of the adaptive learning method, with the addition of a dilate and an erode morphological operations. The goal of using the morphological operations was to reduce common failures (present in the baseline method), and to preserve the original walker height, as illustrated in Fig. 4. Morphological operations also cause orientation and size distortions in the walkers' major axes. Also, instead of comparing pixels in gray level, moving objects and background are compared across the three RGB image channels.

To assess the performance of our method, a special data set with 2000 frames, gathered from PETS 2006 [13], was used with real surveillance scenes, four different camera positions and orientations, (including walkers' occlusion) and adaptive background. Only moving objects were considered to be annotated as the ground truth. The best four public available methods evaluated in [14] were compared against ours: Standard adaptive background learning, mixture of gaussian [15], multi-layer [16] and adaptive SOM [17].

Two metrics were considered to compare the methods: frames per seconds (FPS) and the area under receiver operating characteristic curve (AUC) as shown in Fig. 3. The standard adaptive background learning was the fastest among all, however with the worst AUC. Multi-layer was the slowest one. Our method presented the best AUC, and it was near the fastest method, becoming the most feasible method to be applied in our approach.

IV. Vanishing Point Detection

After suppressing moving objects with the BS method, line segments are found by a modified version of the line segment detector (LSD) [18], as illustrated in Figs. 6 (Left column), at different camera locations. These line segments are the scene structure clues, and they were extracted from static structures presented in the background image. This bundle of line segments allows the use of a reliable and specialized method to estimating vanishing points. Based on Manhattan World assumption, these lines segments point to, at least, 3 vanishing points, which is enough for camera calibration. This approach does not need major axes in the estimation of vanishing points. This increases vanishing point accuracy, allowing for the use of few samples for calibration parameter convergence.

After line segment detection, a RANSAC is used to find the vanishing points with four examples of line segments, which is assumed to follow a Manhattan World assumption. Prior to the RANSAC computation, our method makes an oriented search for the samples used in RANSAC. During the line segment extraction, a histogram of line segment orientations (between 0° and 180°) is built, searching for convergence peaks. Line segments inside the regions of the peaks in the histogram demand few RANSAC iterations to converge to the vanishing points.

Although there are other methods to detect the vanishing points from line segments, our method demonstrated to be the fastest found, detecting enough vanishing points necessary for camera calibration, as well as, avoiding analyzing all points; some results are shown in Figs. 6(Right column). Vanishing points allow defining some camera parameters (f, θ and ρ), and help selecting the walkers' major axes by using the convergence filter (good walkers). This latter chooses line segments by means of Liebowitz's distance [11] to the vertical vanishing points, as will be described further.

The convergence filter selects the principal axis which point to the vertical vanishing point, restricting blobs with wrong orientation. Yet simple, the convergence filter eliminates approximately 90% of the noise presented in the set of principal axis. This effect is illustrated in Figs. 5(a) and (b). With a set of more consistent and smaller amount of data, the reference height procedure (see Sec.) requires less iteration during the selection of the best sampling set.

V. Walkers' Height Distribution

After determining vanishing points, focal length, f, was found according Eq. 2, while ρ and θ were determined by Eqs. 4 and 3, respectively. So far, the only missing parameter is the camera height. To keep the premisse to provide a fully automatic camera calibration framework, the camera height, h_c, is estimated by approximating a height distribution relative to real height distribution. According to Liu et al. [9], human real height distribution encompasses 90% of the heights, h_i^{3D}, around the average height, H^{3D}. Yet, the set of human real heights has a relative difference of less than 7.6% from the mean. This prior distribution allows matching the average human height to the mean of the walkers' relative heights (as shown in Eq. (6)), and it is given by

$$\frac{|h_i - \mu|}{\mu} \leq 0.076 \tag{7}$$

where μ is the mean value of the relative heights.

To fit the relative height to the major' axes found, a RANSAC is used to find the best samples which define the distribution. Randomly, RANSAC selects samples to compute average relative height. After that, Eq. (7) is applied to each sample. The sample is discarded in case of overcoming the boundary defined in Eq. (7). This process is repeated iteratively, and the biggest sample is returned, at the end. After that, the mean, μ, of the relative heights its standard deviation

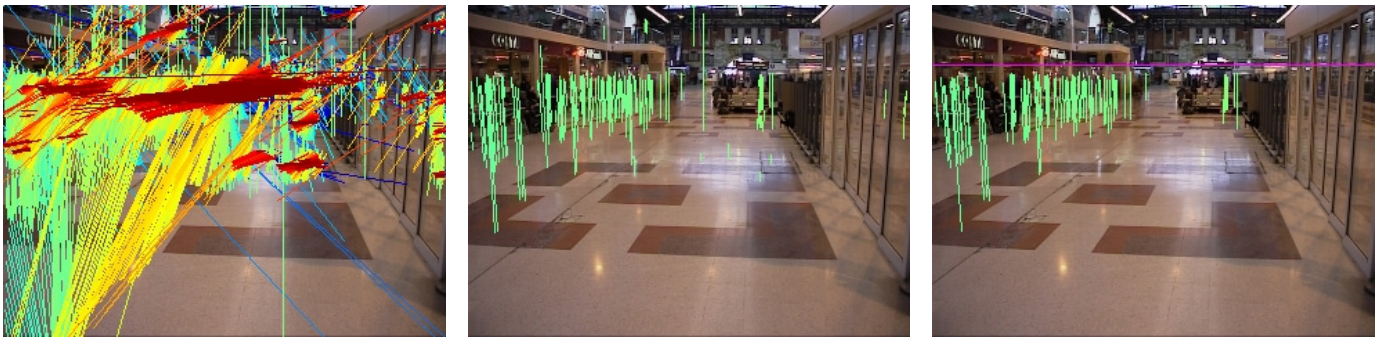

Fig. 5. (a) All extracted major axes; (b) major axes after the convergence filter; (c) selected walkers' major axes which fit to a human height distribution.

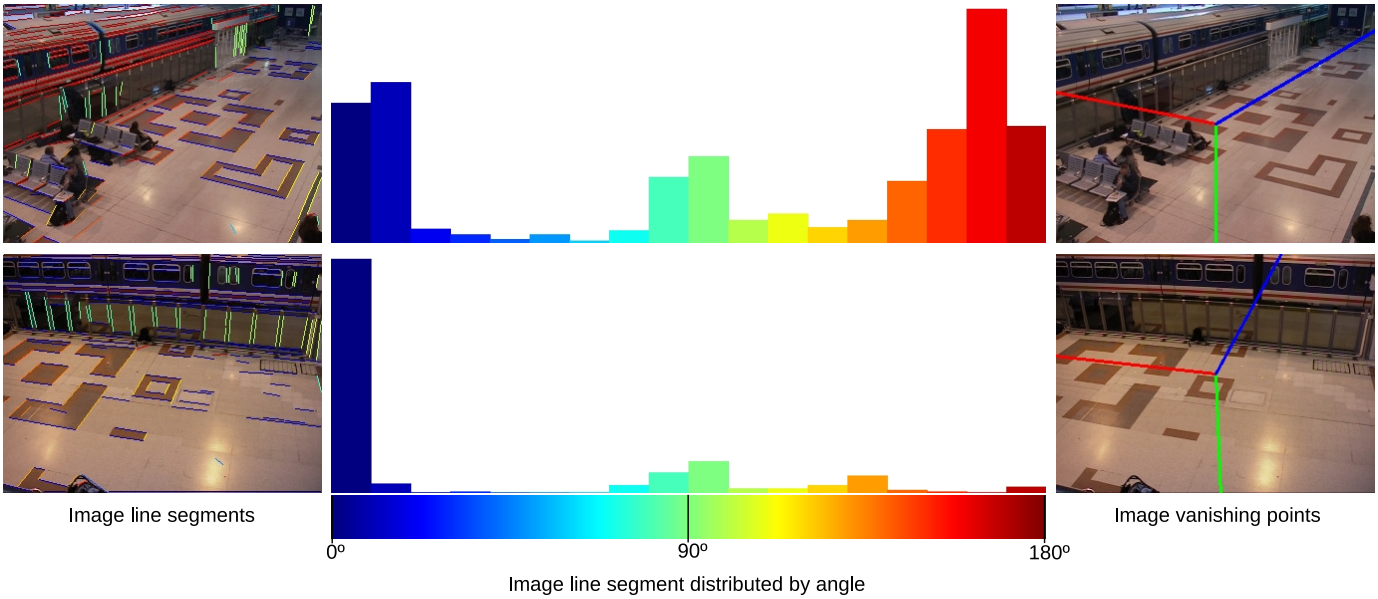

Fig. 6. Examples of vanishing point detection. Left column shows detections of line segments with LSD method [18]; right column show three detected vanishing points.

are calculated. The heights form a normal distribution, which allow to restrict the search to 95% of the elements in 2σ from the mean. This constraint is applied to 90% of the elements (relative heights), providing the selection of the relative walkers' heights by means of the distance between each height and the mean. To each RANSAC iteration, a mean value of the walkers' relative height is assigned considering the interval $[\mu - 2\sigma, \mu + 2\sigma]$.

VI. Experimental Evaluation

Evaluation of the proposed method was assessed from four different data sets: PETS 2006 [13], PETS 2007 [19], CVLAB [20], and a private gathered data set. The goal was to evaluate the value of the relative foot-head homology root mean square error (FHH RMSE), ground-plane measurements (determined by the distance between two points in the ground plane) and walkers' real heights. Each data set provided a way to analyse those metrics, as showed in Table II, which summarizes the characteristics of each one of the data sets, regarding the type of environment (indoor/outdoor), occupation of the scene by the people (partially crowded, crowded or not crowded), how the scene was captured concerning the randomness of people walking (uncontrolled, partially controlled or controlled), number of views and frames evaluated, and the applied type of error analysis. Additionally, the number of necessary good walkers was computed in order to assess how fast the method find the camera parameters. To the best of our knowledge is the first time that a thorough performance assessment of this kind is made.

Figure 7 shows the evaluation of the heights and the distance of points on the ground plane with respect to the number of samples tested, over three different camera heights (close to the floor, around 2 meters, and around 4 meters). Each video takes less than 1 minute to process the scene at 25 FPS, and all walkers' heights are previously known. Height evaluation

TABLE II
CHARACTERISTICS OF THE DATA SETS USED

Dataset	Enviroment	Conditions	Scene	# of views	# of frames	Type of analysis
PETS 2006	indoor	partially crowded	uncontrolled	3	2500	ground-plane measurements FHH RMSE
PETS 2007	indoor	crowded	uncontrolled	2	2500	FHH RMSE
CVLAB	indoor, outdoor	not crowded	partially controlled	5	2500	FHH RMSE
Ours	indoor	not crowded	controlled	3	2500	real height

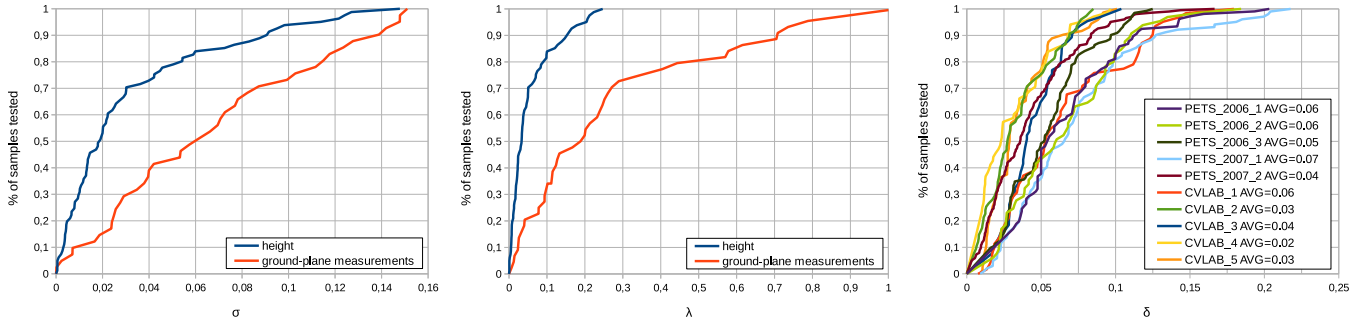

Fig. 7. Cumulative curves of (from left to right): heights, ground-plane measurements and foot-head homology (FHH) RMSE. Left plot: σ represents an upper bound of a relative error found on a percentage of samples tested in the data set; center plot: λ represents an upper bound of an absolute error (in meters) found on the percentage of samples tested in the data set; right plot: δ represents an upper bound of a relative error found on FHH RMSE of samples tested on PETS 2006, PETS 2007 and CVLAB data sets. Our gathered and PETS 2006 data sets were used to evaluate the real heights and ground-plane measurements, respectively.

Fig. 8. Examples of camera auto-calibration results. Green line segments represent the ground truth manually labeled; Red line segments represent the estimated feet-head projection; Cyan circles represent the person position at the ground plane. The first row contains images from PETS 2006 (the first three) and from PETS 2007 (the last two). Second row contains images from CVLAB data set.

demonstrated that our method is able to have an absolute error below 5 cm in 70% of the samples tested (see Fig. 7 (center)). Yet, if one considers, for example, the same 70% of the samples tested, our method presents only 3% of relative error (see Fig. 7 (left)).

The ground truth landmarks in PETS 2006 data sets show the distances between several pair of points on the ground plane. It was observed that distances closer to the camera are more accurate, due to the camera perspective distortion (see Fig. 9). Despite this issue, our method was able to have a maximum relative error of 15% among the estimated distance on the ground (see Fig. 7 (left)). Figure 8 shows visual examples of the achieved results on the PETS 2006 and CVLAB data sets.

The relative FFH RMSE provides a way to evaluate camera calibration accuracy from perspective of image projection, numerically. Our method was submitted to different point of views in CVLAB data sets, and achieved an FHH RMSE of 0.03, overcoming the method proposed by [9], as shown in Table III. These results were achieved even with our proposed method exposed to varying lighting conditions in indoor and outdoor environments. PETS 2006 and 2007 portray data sets with uncontrolled scene and crowded occupation. These data sets showed the method is sensitive to crowd occupation, due the walkers' major axis distortion, which occurred in the foreground segmentation. However, FHH RMSE evaluation

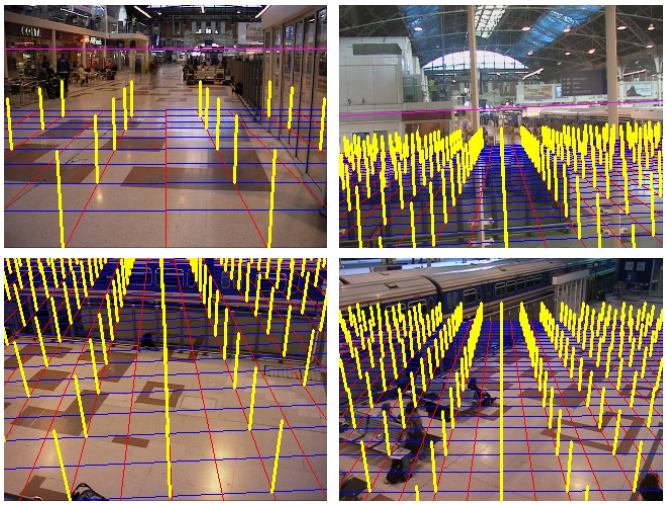

Fig. 9. Ground plane estimation samples for distinct views from PETS 2016. The blue and red lines portray the ground plane, and yellow line segments represent are ground plane perpendicular.

TABLE III
COMPARATIVE EVALUATION OF THE CALIBRATION PROCESS ON THE CVLAB DATA SET.

Metrics	Liu et al [9]	Our Method
Average of good walkers necessary for convergence	1800	370
FHH RMSE	0.05	0.03

reaches 0.07 as a maximum error, indicating that the method is also robust to handle with crowded scene. Table III shows also the comparison of our method against [9] with respect to the average number of good walkers necessary to camera parameter convergence; in this case, our method requires almost 5 times less examples than the method proposed in [9]. Figure 8 shows visual examples of the achieved results on the PETS 2006, PETS 2007, CVLAB data sets.

Our method can handle with a low resolution representation of a person, as shown in Fig. 6 (first row and column), due to the BS method sensitivity. The proposed framework can be applied in several indoor public spaces (e.g. airports, malls, buildings), and outdoor places, mainly occupied by people and with some man-made structures in the background (e.g. squares, parks).

VII. CONCLUSION

In this paper, a novel framework for fully automatic calibration of static surveillance cameras based on scene cues was presented. The proposed method demonstrated to have high accuracy, very fast calibration parameter convergence and no need of human intervention. Although BS methods usually produce a lot of noise in blob extraction – used to define walkers' major axes – the proposed camera calibration framework is able to overcome this problem, still demonstrating high accuracy at the end. The prior information on walkers, along with the walkers' relative heights and scene cues, avoid the need of knowing any reference length in the scene. As future work, we are investigating a method to compute automatically the radial distortion of the camera lens.

REFERENCES

[1] S. Rujikietgumjorn and R. Collins, "Optimized pedestrian detection for multiple and occluded people," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3690–3697.

[2] A. Balan, L. Sigal, and M. Black, "A quantitative evaluation of video-based 3d person tracking," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 349–356.

[3] Z. Wu, Y. Li, and R. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1095–1108, 2015.

[4] F. Lv, Z. T., and R. Nevatia, "Self-calibration of a camera from video of a walking human," in *IEEE International Conference on Pattern Recognition*, 2002, pp. 562–567 vol.1.

[5] N. Krahnstoever and P. Mendonca, "Bayesian autocalibration for surveillance," in *IEEE International Conference on Computer Vision*, 2005, pp. 1858–1865.

[6] F. Lv, T. Zhao, and R. Nevatia, "Camera calibration from video of a walking human," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1513–1518, 2006.

[7] D. Rother, K. Patwardhan, and G. Sapiro, "What can casual walkers tell us about a 3d scene?" in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[8] B. Micusik and T. Pajdla, "Simultaneous surveillance camera calibration and foot-head homology estimation from human detections," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1562–1569.

[9] J. Liu, R. Collins, and Y. Liu, "Surveillance camera autocalibration based on pedestrian height distributions," in *British Machine Vision Conference*, 2011, pp. 144–154.

[10] S. Lee and R. Nevatia, "Robust camera calibration tool for video surveillance camera in urban environment," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2011, pp. 62–67.

[11] D. Liebowitz and A. Zisserman, "Combining scene and auto-calibration constraints," in *IEEE International Conference on Computer Vision.*, 1999, pp. 293–300.

[12] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[13] I. C. on Computer Vision and P. Recognition, "Benchmark data," 2006, available from http://www.cvg.reading.ac.uk/PETS2006/data.html.

[14] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, 2014.

[15] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Springer US, 2002, pp. 135–144.

[16] J. Yao and J. Odobez, "Multi-layer background subtraction based on color and texture," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[17] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on Image Processing*, pp. 1168–1177, 2008.

[18] R. Gioi, J. Jakubowicz, J. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 722–732, 2010.

[19] I. C. on Computer Vision and P. Recognition, "Benchmark data," 2007, available from http://www.cvg.reading.ac.uk/PETS2007/data.html.

[20] C. V. Laboratory, "Multi-camera pedestrians video," 2013, available from http://cvlab.epfl.ch/data/pom/.